Towards the Discovery of Diseases Related by Genes Using Vertex Similarity Measures

Hung-Hsuan Chen[†], Liang Gou[♮], Xiaolong (Luke) Zhang[‡], C. Lee Giles[†‡]
[†]Computer Science and Engineering, [‡]Information Sciences and Technology
The Pennsylvania State University, University Park, PA 16802, USA
[♮]Almaden Research Center, IBM, San Jose, CA 95120, USA
hhchen@psu.edu, lgou@us.ibm.com, {lzhang, giles}@ist.psu.edu

***Abstract*—Discovering the relationships of gene to gene, gene to its related diseases, and diseases implicated in common genes is important. However, traditional biological methods can be expensive. Here, we show that the diseases implicated in common genes and the genes related to a multiple-gene disease can be inferred by the vertex similarity measures, a type of method to find the similar vertices in a network based on its structure. The relationship among diseases and the relationship among genes are modeled as two biological networks: human disease network and disease gene network. We apply the vertex similarity among the vertices in the human disease network to infer the diseases implicated in common genes. By similar manner, we utilize vertex similarity measures on the disease gene network to infer the genes related to a common multiple-gene disease. Experimental results demonstrate the potential of vertex similarity as an inexpensive approach to infer the possible links between genes and between diseases. We also develop a system to visualize and get a better understanding about the relationships among diseases and genes.**

I. Introduction

Discovering 1) the common genes of different genetic diseases and 2) the genes related to a multiple gene disease helps improve the diagnosis and the development of new therapies, and, hopefully, the ability to further understand genetic diseases. For example, autism spectrum disorders (ASDs) and epilepsy have recently been found to have a common predisposing gene; implying that there could be an underlying pathogenesis between autism and epilepsy [8]. Traditional biological methods to locate the disease genes include positional cloning [23], DNA marker [14], and positional candidate approaches [6]. However, these methods usually require manual resources and expensive experiments.

In this paper, we show the potential of inferring unknown relationships between genes and between diseases using known relationships among them. Specifically, the known relationship among genes and diseases is modeled as a bipartite graph consisting of two disjoint sets of vertices: the human genetic diseases set and the genes set. A disease and a gene are connected if the mutation on the gene is implicated in the disease [11]. The bipartite network can be projected to two networks: the human disease network and the disease gene network. The unknown relationships among diseases and genes are suggested by applying vertex similarity measures to the two networks respectively. The vertex similarity measures are one type of calculation that generates similarity scores among vertices in a network based on its structure. We introduce the common local structure based vertex similarity measures (Jaccard similarity [19] and preferential attachment [2, 16]), global structure based similarity (SimRank [13]), and our proposed relation strength similarity (RSS) [4, 5] and compare their ability to predict possible missing links which represent unknown relationships. Experimental results show that on average the precision for link prediction is high. This demonstrates the vertex similarity measures can be an inexpensive indicator for genetic disease relationships and gene relationship analysis. The predicted links among diseases and among genes are shown to be promising candidates that are worth further exploration. We also visualize the relationships and the recommendation links among diseases and among genes.

II. Related Works

Community detection can be useful for many types of social networks. One popular research direction of community detection is based on the intuition that the intra-community edges should be denser than the inter-community edges. A well known example is the max-flow min-cut theorem [9], which determines the minimal cuts to divide the graph. Several studies utilized similar measures such as the ratio cut and normalized cut [12, 21]. However, these methods suffer limitations in resolution. One can only claim the nodes of the same community are similar to each other, but it is difficult to determine the top similar node pairs. Vertex similarity [15] is a decent solution to the problem. It defines the similarity between any pair of nodes based on the referenced property, such as the structure of the network. Thus, the node-to-node level granularity is specifically defined, and the communities are grouped by the similar nodes.

Several biological systems can be represented by a complex network, such as the protein network [20], Genome network [3], and Diseasome network [11]. The network analysis techniques and tools can be naturally applied on such systems. For example, it is shown that the degree distribution of metabolic and protein networks follow power-law [22]. Thus, these networks are usually very robust to random failure of nodes. Several topological features, such as betweenness and degree, were used to infer the importance of a node. Com-

TABLE I
THE STATISTICAL PROPERTIES OF DISEASOME, HDN, DGN, AND THEIR CORRESPONDING GIANT COMPONENTS (GCS). ($|V|$: THE NUMBER OF VERTICES; $|E|$: THE NUMBER OF EDGES; d: THE AVERAGE DEGREE; C: THE AVERAGE CLUSTERING COEFFICIENT; s: THE NUMBER OF SINGLETONS; ℓ: THE AVERAGE SHORTEST PATH LENGTH)

Network	$\|V\|$	$\|E\|$	d	C	s	ℓ
Diseasome	3,061	2,673	1.75	0	0	-
Diseasome GC	1,419	1,550	2.18	0	0	12.29
HDN	1,284	1,527	3.52	0.56	417	-
HDN GC	516	1,188	4.60	0.64	0	6.51
DGN	1,777	7,491	10.87	0.77	399	-
DGN GC	903	6,760	14.97	0.85	0	5.93

munity detection algorithm was used on biological networks to find the similar nodes. For example, food web of marine organisms is successfully divided into pelagic organisms and benthic organisms in [10]. Instead of focusing on each small biological component, the network analysis approaches help us understanding the interaction and relationship between the components in a global view.

III. METHODOLOGY

A. Data Description

The known relationships between genes and diseases are constructed as the Diseasome[1], a bipartite graph with two disjoint sets of vertices [11]. One set contains all known genetic diseases, and the other set includes all known disease genes in the human genome. A disease and a gene are connected if a mutation of the gene would cause the disease.

The Diseasome can be projected to two biologically related networks: a human disease network (HDN) and a disease gene network (DGN). For HDN, each vertex represents a disease. An edge attaches two vertices if there are one or more genes that are implicated in both. Edge weights correspond to the number of common genes between the two diseases. For DGN, each vertex means a disease gene. Two genes are connected if they are associated with at least one common disease. The edge weights signify the number of diseases with which the two genes are mutually associated.

The statistical properties of Diseasome, HDN, DGN, and their corresponding giant components are shown in Table I. The average clustering coefficients for Diseasome and its giant component are 0 because they are bipartite graph thus the neighbors of a given vertex are never connected to each other. The average shortest path length and the diameter are not shown in Diseasome, HDN, and DGN because none of them are connected graphs.

B. Vertex Similarity Measures

HDN and DGN provide the disease-centered and gene-centered view of Diseasome respectively. In both networks, two vertices are connected if they are related. Since we are interested in the vertices that are related but still unknown, the problem can be modeled as a missing link prediction problem, which aims to discover the potential links in the network.

[1]http://diseasome.eu/

TABLE II
TIME COMPLEXITY COMPARISON OF JACCARD, PREFERENTIAL ATTACHMENT, SIMRANK, AND RSS. (n: NUMBER OF VERTICES; d: AVERAGE DEGREE; K: MAXIMUM NUMBER OF ITERATIONS; r: DISCOVERY RANGE, ASSUMING $d << n, K << n, r << n$)

Vertex Similarity Measure	Time Complexity
Jaccard	$O(nd^3) \sim O(n)$
preferential attachment	$O(n^2d^2) \sim O(n^2)$
SimRank	$O(Kn^2d^2) \sim O(n^2)$
RSS	$O(nd^r) \sim O(n)$

The missing link can be inferred by the intrinsic properties of the vertices, such as the symptom of the diseases in HDN or the phenotypes of the genes in DGN. However, measuring the intrinsic properties of the vertices in biological network usually requires huge human power, expensive experiments, and abundant domain knowledge about the target. The other type of missing link prediction approach is vertex similarity based measures, which determines the missing links based on the similarity among vertices of a network using the network structure. For networks with many known links and several links have not yet been observed, the vertex similarity measure is a cheap tool to capture the possible links [4, 5, 17].

We apply several vertex similarity measures on HDN and DGN, the two projections of Diseasome, to show the potential of vertex similarity measures for missing link prediction in biological related networks. The vertex similarity measures used in the paper include two local structure based measures (Jaccard similarity [19] and preferential attachment [2, 16]), one global structure based similarity (SimRank) [13], and our proposed relation strength similarity [4, 5], which has a parameter r to control the discovering range, i.e., the number of hops away for each vertex to explore the missing links.

Jaccard similarity measure is based on the intuition that two vertices are more similar if they share more common neighbors. Studies show that it is good at predicting the missing links in coauthorship network [4, 5, 17]. For two vertices v_i and v_j, Jaccard similarity is defined by Equation 1.

$$S_{Jaccard/cosine}(v_i, v_j) := \frac{|m_i \cap m_j|}{|m_i \cup m_j|}, \tag{1}$$

where m_i is the set of neighbors of vertex v_i, the $|\cdot|$ function returns the number of elements in the set.

The preferential attachment is based on the observation that a high degree node is more likely to acquire new links. The phenomenon is common in several large scale networks, such as World Wide Web [1], citation network [18], and protein network [7]. Newman [16] proposed that the probability of a new edge established between two vertices is proportional to the product of their degree, as defined in Equation 2.

$$S_{pref-attach}(v_i, v_j) := |m_i| \cdot |m_j|. \tag{2}$$

Instead of using the topology information near the given vertices, SimRank considers the global structure of the network by a recursive definition: two vertices are similar if their direct

neighbors are themselves similar, as defined in Equation 3.

$$S_{SimRank}(v_i, v_j) := c\frac{\sum_{\forall x \in m_i} \sum_{\forall y \in m_j} S_{SimRank}(v_x, v_y)}{|m_i| \cdot |m_j|}, \tag{3}$$

where c is a parameter specifying the relative importance ratio between the indirect neighbors and direct neighbors for similarity calculation ($0 \leq c \leq 1$). The smaller the value of c, the less important the indirect neighbors are.

The other vertex similarity measure, relation strength similarity, permits users to explicitly assign the weights to every edge proportional to the relation strength between vertices. In HDN, the edge weights are the number of known common genes related to two diseases; whereas in DGN, the edge weights represent the number of known common diseases related to two genes. RSS [4, 5] considers the path length, the number of distinct simple paths, and the relation strength between neighboring vertices for similarity calculation. Equation 4 defines RSS.

$$S_{RSS}(v_i, v_j) := \sum_{m=1}^{M} R^*_{p_m}(v_i, v_j), \tag{4}$$

where $R^*_{p_m}(v_i, v_j)$ is the general relation strength from v_i to v_j along the path p_m, which is defined in Equation 5.

$$R^*_{p_m}(v_i, v_j) := \begin{cases} \prod_{k=1}^{K} R(u_k, u_{k+1}) & \text{if } K \leq r \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where r is the discovery range parameter controlling the maximum degree of separation to calculate, and p_m is a path from $v_i (= u_1)$ to $v_j (= u_K)$ through vertices $u_2, u_3, \ldots, u_{K-1}$.

The term $R(u_k, u_{k+1})$ in Equation 5 is relation strength (or the normalized edge weight), as defined in Equation 6.

$$R(v_i, v_j) := \begin{cases} w_{ij} / \sum_{\forall k \in m_i} w_{ik} & \text{if } v_i \text{ and } v_j \text{ are neighbors} \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

where w_{ij} is the weight for the edge (v_i, v_j).

One advantage of RSS over previous introduced vertex similarity measures is that RSS allows users to specifically assign weights to the edges. A higher edge weight infers the two end nodes have a stronger connection or more frequent interaction; thus they should have a higher similarity score.

Table II lists the time complexity to compute these vertex similarity measures.

IV. Experiments

In this section, we show the potential of the vertex similarity measures in terms of their ability to predict the missing links in the two biological networks HDN and DGN.

A. Experiment Setup

One way to determine the validity of vertex similarity measures in discovering unknown relationship for HDN and DGN is biological experiments such as DNA marker [14] and positional candidate approach [6]. However, these approaches usually require expensive experiments.

Instead of conducting the expensive biological experiments to verify the results, we imitate the machine learning technique by separating the known information into training and testing data set to show the potential of vertex similarity measures [5]. Specifically, for the $1,527$ known links in the HDN, each link has a probability p to be included in the training network and $(1-p)$ in the testing network ($0 < p < 1$). The expected numbers of links in the training network and testing network are $1,527p$ and $1,527(1-p)$ respectively. We perform the same setup for DGN. In addition, among the $1,284$ vertices in HDN, 417 of them are singletons, i.e., the vertices have no links attached to it. The singletons are removed because the similarity score between a singleton and any other vertices is always zero by vertex similarity measures. Thus, the training network of HDN contains 867 vertices. By similar manner, the training network of DGN has $1,378$ vertices.

We apply the vertex similarity measures on the training network to obtain the similarity scores of each non-neighbor vertex pair. The potential links are predicted by requiring the top-n most similar pairs be connected. The correctness of the prediction is validated by the testing network. The procedure is repeated 20 times independently.

B. Experimental Results

To evaluate the prediction performance, a commonly used measure is $Prec(S_m, n)$ (precision at n), which gives a cut-off rank of precision by considering only the topmost results.

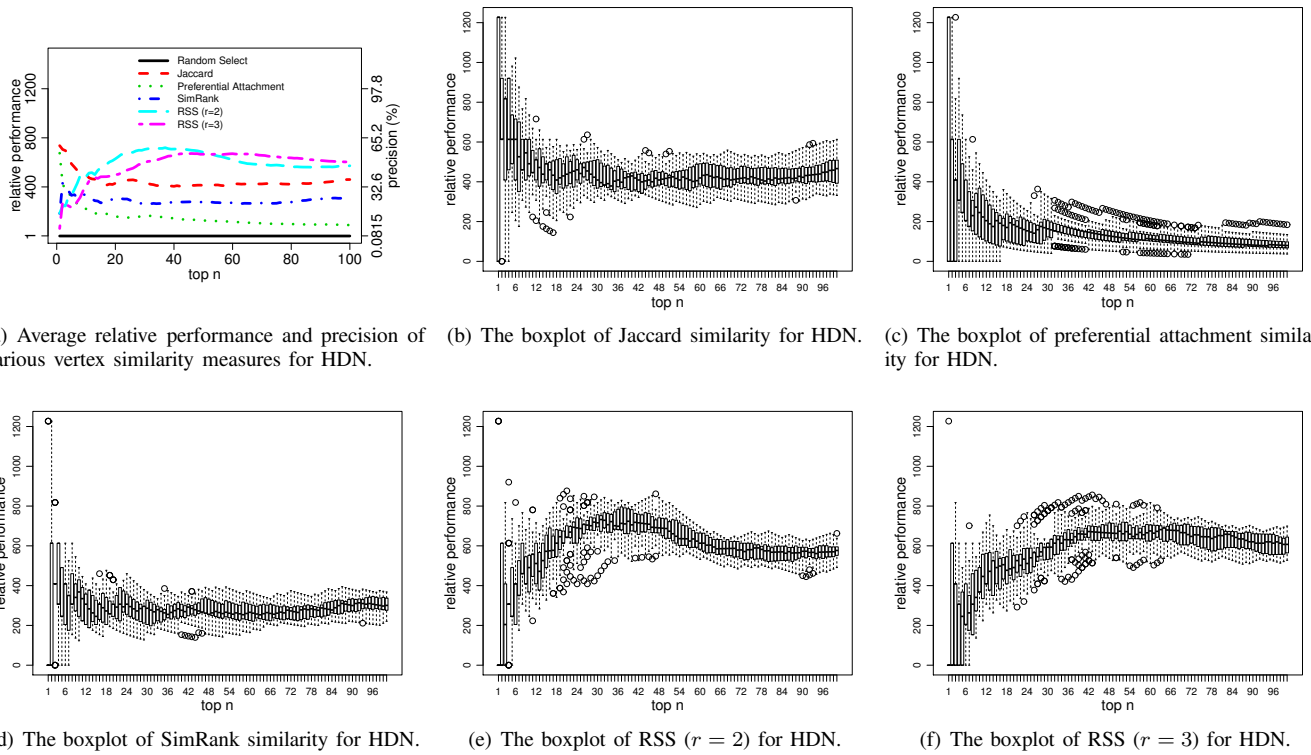
Unlike the coin flip guessing problem which has 50% precision by naïve random guessing, link prediction is much harder because the precision of a random guess is very low [5, 17]. When the training network contains $p = 80\%$ of the edges of the original network, the training network of HDN would have 867 vertices and $1,222$ edges. Randomly picking two vertices and requiring the two should be connected gives $\binom{867}{2} = 375,411$ possible combinations. Since $1,222$ of them are already connected in the training network, there are $375,411 - 1,222 = 374,189$ non-neighbor pairs. Only $1,527(1-p) = 305$ of them are the correct pairs. Thus, the precision of a naïve random pick for the HDN is $305/374,189 = 0.0815\%$. By similar calculation, the precision of a random pick for DGN is 0.0953%.

In Figure 1(a), the average precision of 20 independent trials for different vertex similarity measures in HDN is shown. To demonstrate the effectiveness of vertex similarity measures, for each measure we show both the precision and relative performance $P(S_m, n)$, which is defined in Equation 7.

$$P(S_m, n) := \frac{Prec(S_m, n)}{Prec(S_r, n)}, \tag{7}$$

where S_m is the given vertex similarity measure, S_r is a naïve random select measure, $Prec(S_m, n)$ is the precision of S_m by requiring the top-n similar vertex pairs should be connected. A larger relative performance score is preferred.

As shown, Jaccard similarity is good when n is smaller than 10. When n is between 11 and 100, RSS outperforms all other measures for both $r = 2$ and $r = 3$. Although



(a) Average relative performance and precision of various vertex similarity measures for HDN.

(b) The boxplot of Jaccard similarity for HDN.

(c) The boxplot of preferential attachment similarity for HDN.

(d) The boxplot of SimRank similarity for HDN.

(e) The boxplot of RSS ($r = 2$) for HDN.

(f) The boxplot of RSS ($r = 3$) for HDN.

Fig. 1. The relative performance ratio of various vertex similarity measures for HDN.

SimRank considers the global structure, its performance is not as good as the local structure based Jaccard similarity and our proposed RSS. The preferential attachment is good at simulating the global network statistics [2]. However, it is less effective in terms of the ability to predict individual missing links. Even the worst preferential attachment measure is more than 147.95 times better than random select in average. This demonstrates the potential of vertex similarity measures as the non-expensive indicators for the genetic diseases sharing common genes. To let the readers see more insight about the experimental results, Figure 1(b) to Figure 1(f) show the box-and-whisker plot for the 20 independent experiments of Jaccard similarity, preferential attachment similarity, SimRank, RSS with discovery range 2, and RSS with discovery range 3. They suggest that the vertex similarity measures are very stable in predicting the missing links of HDN.

By the same manner, the average performance of all the vertex similarity measures on DGN is shown in Figure 2(a). The Jaccard similarity performs best in DGN, followed by our proposed RSS. SimRank is a little behind RSS. Preferential attachment is again the worst among all the similarity measures. Except preferential attachment, all the similarity measures are more than 500 times better than random select in average. Figure 2(b) to Figure 2(f) show the box-and-whisker plot of these measures. They suggest that vertex similarity measures are superior and stable indicator to identify the genes that are related to a common multi-gene disease.

V. Missing Link Prediction

The previous section separate the known information into training and testing network to demonstrate the potential vertex similarity measures on biological networks, such as HDN and DGN. In this section, we show the missing links predicted by various vertex similarity measures using the full topology of the given networks. A visualization system is also introduced.

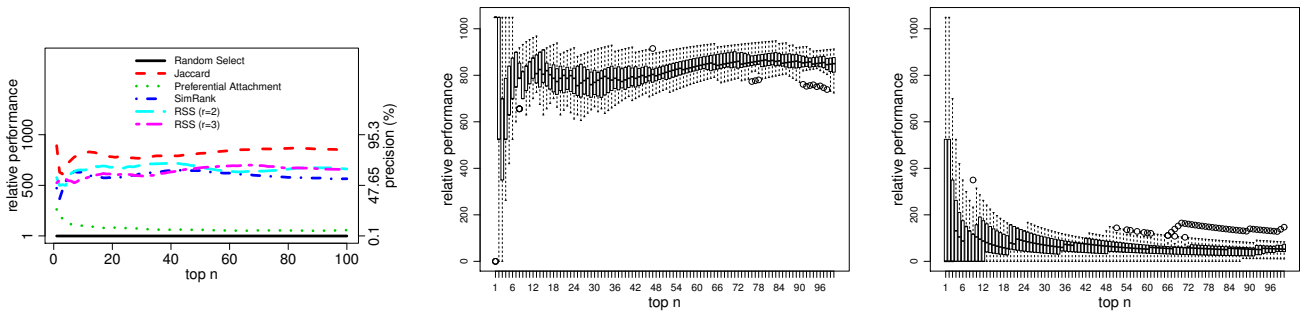
A. Link Prediction Results

By using all the known relationship, Table III and Table IV present the predictions of the top 5 similar vertices by using different vertex similarity measures on HDN and DGN respectively. The relationships have not yet been validated by biological experiments, but they could be the potential disease pairs or gene pairs that deserve more attention. One of our ongoing work is working with biologists for validation.

B. Link Prediction Visualization

The Tables introduced in previous section give only the predicted results. To get a deeper understanding about how the two non-neighboring diseases or genes are related, we implemented a system to visualize the results. Users can interact with the system to see the relationship between diseases or between genes. The results predicted by different similarity measures can also be compared and visualized.

Figure 3(a) is a snapshot of HDN in the demonstration program. The right hand side shows the disease network in which each vertex is a human disease, an edge represents at least one known gene is implicated in two diseases. The disease names appear when moving the cursor on the vertices.



(a) Average relative performance and precision of various vertex similarity measures for DGN.

(b) The boxplot of Jaccard similarity for DGN.

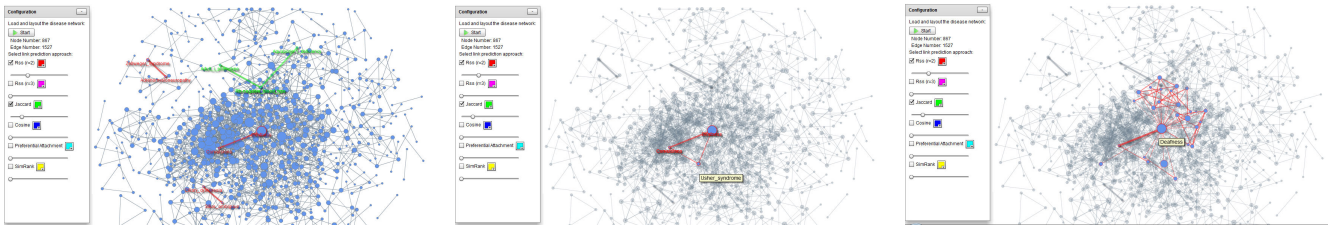
(c) The boxplot of preferential attachment similarity for DGN.

(d) The boxplot of SimRank similarity for DGN.

(e) The boxplot of RSS ($r = 2$) for DGN.

(f) The boxplot of RSS ($r = 3$) for DGN.

Fig. 2. The relative performance ratio of various vertex similarity measures for DGN.



(a) A snapshot of HDN in the demonstration program.

(b) A snapshot of how to diseases are related.

(c) A snapshot of the disease of interest and its neighbors

Fig. 3. Snapshots of the visualization system.

The sizes of the vertices are proportional to their degrees. The left panel lists available measures: RSS with $r = 2$, RSS with $r = 3$, Jaccard similarity, preferential attachment, and SimRank. Users can choose from these measures for potential link recommendation. The slide bar under each measure is used to specify the number of recommending links to show. In the snapshot, RSS with $r = 2$ and Jaccard similarity are checked. The top 3 similar non-neighbor vertices calculated by RSS with $r = 2$ and the top 2 similar non-neighbor vertices calculated by Jaccard similarity are shown by red and green lines respectively.

To see how two suggested vertices are related, users can double click on the recommendation link. As Figure 3(b) shows, the mutual neighbor vertices are highlighted to help users understand how and why two vertices might be related.

A user interested in a particular disease can double click on the corresponding vertex to highlight the vertex and its neighbors. As shown in Figure 3(c), the disease "Deafness" and all the known diseases that are implicated with genes in common with Deafness can be highlighted.

The DGN can also be represented in a similar manner. Due to space limitations, we only show the snapshots of HDN.

VI. Conclusions and Future Work

We show that mining relationships among genetic diseases and among genes using vertex similarity measures can be an inexpensive and promising indicator for potential gene-disease relationship discovery. By requiring the most similar vertex pairs should be connected, the average precision for human disease network is about 60%, more than 700 times better than the naïve random selection. The performance on disease gene network is even better: the average precision by requiring the first returned pair should be connected is 85%, almost 900 times better than the random selection. This suggests the effectiveness of using vertex similarity measures to explore the potential links for conducting actual biological experiments.

We list the top 5 similar vertices on human disease network and disease gene network as promising related diseases and related genes respectively. In addition, a visualization system

TABLE III
THE TOP-5 PREDICTED RESULTS OF THE GENETIC DISEASES THAT MAY HAVE GENES IN COMMON.

	RSS ($r = 2$)	RSS ($r = 3$)	Jaccard	Pref. Attach.	SimRank
1	**Situs ambiguus** and **Ciliary dyskinesia**	**Adrenomyeloneuropathy** and **Zellweger syndrome**	**Tolbutamide poor metabolizer** and **Vitamin K-dependent coagulation defect**	**Colon cancer** and **Deafness**	**Tolbutamide poor metabolizer** and **Vitamin K-dependent coagulation defect**
2	**Convulsions** and **Deafness**	**Rhizomelic chondrodysplasia punctata** and **Zellweger syndrome**	**Palmoplantar keratoderma** and **Steatocystoma multiplex**	**Colon cancer** and **Diabetes mellitus**	**Palmoplantar keratoderma** and **Steatocystoma multiplex**
3	**Adrenomyeloneuropathy** and **Zellweger syndrome**	**Situs ambiguus** and **Ciliary dyskinesia**	**Night blindness** and **Retinal cone dsytrophy**	**Colon cancer** and **Prostate cancer**	**Night blindness** and **Retinal cone dsytrophy**
4	**Rhizomelic chondrodysplasia punctata** and **Zellweger syndrome**	**Convulsions** and **Deafness**	**Walker-Warburg syndrome** and **Myotilinopathy**	**Colon cancer** and **Retinitis pigmentosa**	**Walker-Warburg syndrome** and **Myotilinopathy**
5	**Palmoplantar keratoderma** and **Steatocystoma multiplex**	**Hyperproinsuli-nemia** and **Diabetes mellitus**	**Maple syrup urine disease** and **Pyruvate dehydrogenase deficiency**	**Deafness** and **Breast cancer**	**Maple syrup urine disease** and **Pyruvate dehydrogenase deficiency**

TABLE IV
THE TOP-5 PREDICTED RESULTS OF THE GENES THAT MIGHT BE RELATED TO A COMMON MULTI-GENETIC DISEASE.

	RSS ($r = 2$)	RSS ($r = 3$)	Jaccard	Pref. Attach.	SimRank
1	**ZIC3** and **THRAP2**	**AHI1** and **NPHP4**	**SLC25A19** and **MCPH1**	**EYA4** and **TP53**	**ZIC3** and **THRAP2**
2	**TNNI2** and **TPM2**	**SDHC** and **SDHB**	**ZIC3** and **THRAP2**	**KRAS** and **EYA4**	**WWOX** and **TGFBR1**
3	**SDHC** and **SDHB**	**PSORS6** and **BTNL2**	**WWOX** and **TGFBR1**	**TP53** and **PPARG**	**CHX10** and **BCOR**
4	**LZTS1** and **TGFBR2**	**PSORS6** and **HLA-DRB1**	**CHX10** and **BCOR**	**CCND1** and **EYA4**	**OCA2** and **TYRP1**
5	**PPP1R3A** and **PPARG**	**ZIC3** and **THRAP2**	**OCA2** and **TYRP1**	**KRAS** and **PPARG**	**NPC1** and **SMPD1**

has been build and is demonstrated to help visualize and understand the relationships and their predictions.

Future work could be both computational and biological. Of course, it would be important to validate these predictions. Machine learning methods could also be applied to predict missing links and hopefully obtain a higher precision rate. Network algorithms presented here could be applied to other types of biological networks to study their characteristics.

REFERENCES

[1] A. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509, 1999.

[2] A. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3-4):590–614, 2002.

[3] A. Butte and I. Kohane. Creation and implications of a phenome-genome network. *Nature biotechnology*, 24(1):55–62, 2006.

[4] H.-H. Chen, L. Gou, X. Zhang, , and C. L. Giles. Collabseer: A search engine for collaboration discovery. In *Proceedings of the 11th ACM/IEEE-CS Joint Conference on Digital Libraries*. ACM, 2011.

[5] H.-H. Chen, L. Gou, X. Zhang, , and C. L. Giles. Discovering missing links in networks using vertex similarity measures. In *Proceedings of the 27th ACM Symposium On Applied Computing*. ACM, 2012.

[6] F. Collins. Positional cloning moves from perditional to traditional. *Nature Genetics*, 9(4):347–350, 1995.

[7] E. Eisenberg and E. Levanon. Preferential attachment in the protein network evolution. *Physical review letters*, 91(13):138701, 2003.

[8] A. Fassio, L. Patry, S. Congia, F. Onofri, A. Piton, J. Gauthier, D. Pozzi, M. Messa, E. Defranchi, M. Fadda, et al. Syn1 loss-of-function mutations in asd and partial epilepsy cause impaired synaptic function. *Human Molecular Genetics*, 2011.

[9] L. Ford and D. Fulkerson. Maximal flow through a network. *Canadian Journal of Mathematics*, 8(3):399–404, 1956.

[10] M. Girvan and M. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821, 2002.

[11] K. Goh, M. Cusick, D. Valle, B. Childs, M. Vidal, and A. Barabási. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21), 2007.

[12] L. Hagen and A. Kahng. New spectral methods for ratio cut partitioning and clustering. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 11(9):1074–1085, 1992.

[13] G. Jeh and J. Widom. SimRank: A measure of structural-context similarity. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 538–543. ACM, 2002.

[14] B. Kerem, J. Rommens, J. Buchanan, D. Markiewicz, T. Cox, A. Chakravarti, M. Buchwald, and L. Tsui. Identification of the cystic fibrosis gene: genetic analysis. *Science*, 245(4922):1073, 1989.

[15] E. Leicht, P. Holme, and M. Newman. Vertex similarity in networks. *Physical Review E*, 73(2):026120, 2006.

[16] M. Newman. Clustering and preferential attachment in growing networks. *Physical Review E*, 64(2):25102, 2001.

[17] D. L. Nowell and J. Kleinberg. The link prediction problem for social networks. In *CIKM03: Proceedings of the Twelfth International Conference on Information and Knowledge Management*, pages 556–559, 2003.

[18] D. Price. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5), 1976.

[19] G. Salton. *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. 1989.

[20] V. Satuluri, S. Parthasarathy, and D. Ucar. Markov clustering of protein interaction networks with improved balance and scalability. In *Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology*, pages 247–256. ACM, 2010.

[21] J. Shi and J. Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905, 2000.

[22] S. Wuchty, E. Ravasz, and A. Barabási. The architecture of biological networks. *Complex systems science in biomedicine*, pages 165–181, 2006.

[23] Y. Zhang, R. Proenca, M. Maffei, M. Barone, L. Leopold, and J. Friedman. Positional cloning of the mouse obese gene and its human homologue. *Nature*, 372(6505):425, 1994.